NLTK através de exemplos:
- Modelagem de tópicos (*topic modeling*)
- Redes de co-ocorrência

Prof. Jesús P. Mena-Chalco

jesus.mena@ufabc.edu.br

2Q-2019

Seeking Life's Bare (Genetic) Necessities

Genética?

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Genes in common 233 genes

Haemophilus genome 1703 genes

Mycoplasma genome 469 genes

Genes needed for biochemical pathways +22 genes

256 genes

Redundant and parasite-specific genes removed – 4 genes

Minimal gene set 250 genes

Related and modern genes removed –122 genes

128 genes

Ancestral gene set

ADAPTED FROM NCBI

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

2

Genética?

Computação?

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

3

Genética?

Computação?

Biologia?

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Haemophilus genome
1703 genes

Genes in common
233 genes

Mycoplasma genome
469 genes

Genes needed for biochemical pathways +22 genes

256 genes

Redundant and parasite-specific genes removed – 4 genes

Minimal gene set 250 genes

Related and modern genes removed –122 genes

128 genes

Ancestral gene set

ADAPTED FROM NCBI

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

4

Este exemplo evidencia que um artigo/documento pode estar composto de **diferentes tópicos** (ou unidades) que se misturam para a elaboração de ideias.

Genética?

Computação?

Biologia?

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Haemophilus genome 1703 genes

Genes in common 233 genes

Mycoplasma genome 469 genes

Genes needed for biochemical pathways +22 genes

Redundant and parasite-specific genes removed – 4 genes

Related and modern genes removed –122 genes

256 genes

Minimal gene set 250 genes

128 genes

Ancestral gene set

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

Artigo de revisão em Communication of ACM - 2012

Surveying a suite of algorithms that offer a solution to managing large document archives.

BY DAVID M. BLEI

Probabilistic Topic Models

AS OUR COLLECTIVE knowledge continues to be digitized and stored—in the form of news, blogs, Web pages, scientific articles, books, images, sound, video, and social networks—it becomes more difficult to find and discover what we are looking for. We need new computational tools to help organize, search, and understand these vast amounts of information.

Right now, we work with online information using two main tools—search and links. We type keywords into a search engine and find a set of documents related to them. We look at the documents in that set, possibly navigating to other linked documents. This is a powerful way of interacting with our online archive, but something is missing.

Imagine searching and exploring documents based on the themes that run through them. We might "zoom in" and "zoom out" to find specific or broader themes; we might look at how those themes changed through time or how they are connected to each other. Rather than finding documents through keyword search alone, we might first find the theme that we are interested in, and then examine the documents related to that theme.

For example, consider using themes to explore the complete history of the New York Times. At a broad level, some of the themes might correspond to the sections of the newspaper—foreign policy, national affairs, sports. We could zoom in on a theme of interest, such as foreign policy, to reveal various aspects of it—Chinese foreign policy, the conflict in the Middle East, the U.S.'s relationship with Russia. We could then navigate through time to reveal how these specific themes have changed, tracking, for example, the changes in the conflict in the Middle East over the last 50 years. And, in all of this exploration, we would be pointed to the original articles relevant to the themes. The thematic structure would be a new kind of window through which to explore and digest the collection.

But we do not interact with electronic archives in this way. While more and more texts are available online, we simply do not have the human power to read and study them to provide the kind of browsing experience described above. To this end, machine learning researchers have developed *probabilistic topic modeling*, a suite of algorithms that aim to discover and annotate large archives of documents with thematic information. Topic modeling algorithms are statistical methods that analyze the words of the original texts to discover the themes that run through them, how those themes are connected to each other, and how they change over

» **key insights**

■ Topic models are algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents. Topic models can organize the collection according to the discovered themes.

■ Topic modeling algorithms can be applied to massive collections of documents. Recent advances in this field allow us to analyze streaming collections, like you might find from a Web API.

■ Topic modeling algorithms can be adapted to many kinds of data. Among other applications, they have been used to find patterns in genetic data, images, and social networks.

Probabilistic Topic Models | April 2012 | Communications of the ACM

https://m-cacm.acm.org/magazines/2012/4/...probabilistic-topic-models/fulltext?...true ▼

by DM Blei - Cited by 4531 - Related articles

LDA and **probabilistic models**. LDA and other **topic models** are part of the larger field of **probabilistic modeling**. In generative **probabilistic modeling**, we treat our data as arising from a generative process that includes hidden variables.

Ideia: Um documento textual é uma mistura de tópicos.

gene	0.04		data	0.02		life	0.02		brain	0.04
dna	0.02		number	0.02		evolve	0.01		neuron	0.02
genetic	0.01		computer	0.01		organism	0.01		nerve	0.01

Genetics   Computation   Life sciences   Anatomy

Seeking life's bare (genetic) necessities
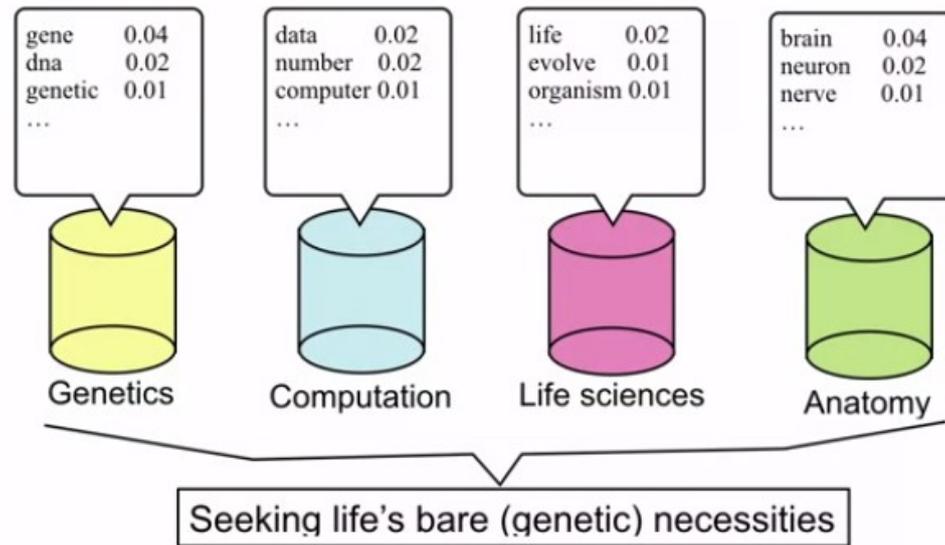
Ao termos um ***novo documento*** ao ser analisado, e considerando apenas esses quatro tópicos, podemos pensar que para esse documento, o mais provável é que seja de **computação**.

Novo documento
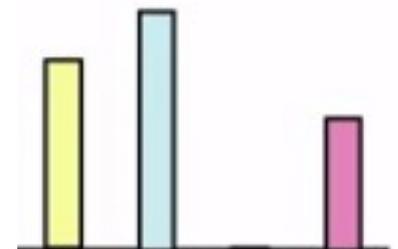
O que é um tópico?

Um tópico pode ser entendido como: Um tema (ou assunto) de um discurso.

Os tópicos podem ser representados por uma distribuição de palavras. Isso significa que uma palavra tem certa probabilidade de aparecer nesse tópico.

Exemplo de tópicos

Tópicos
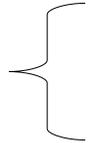
"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Modelagem de tópicos (MT)?

É uma análise "abstrata" (em alto nível) do conteúdo de documentos.

É apropriado quando, frente a um **corpus grande**, deseja-se entender, de forma rápida, o conteúdo dele:

- Trata-se de um documento de "esportes"?
- Trata-se de um documento de "genética"?
- Trata-se de um documento de "computação"?

A **MT** também pode ser utilizada para identificar: (i) como esses tópicos se conectam, e (ii) como mudam ao longo do tempo.

Modelagem de tópicos (MT)

O que geralmente é considerado como entrada para a MT são:

- Uma **coleção de documentos** (ou corpus).
- **Número** finito de tópicos.

O que não sabemos:

- Os **nomes** dos tópicos.

  Não podemos informar, por exemplo, a busca de tópicos em computação.

- A **distribuição de tópicos** para cada documento.

  Não sabemos se os documentos tem um conteúdo assodiado a 50% computação e 50% esportes.

Modelagem de tópicos (MT)
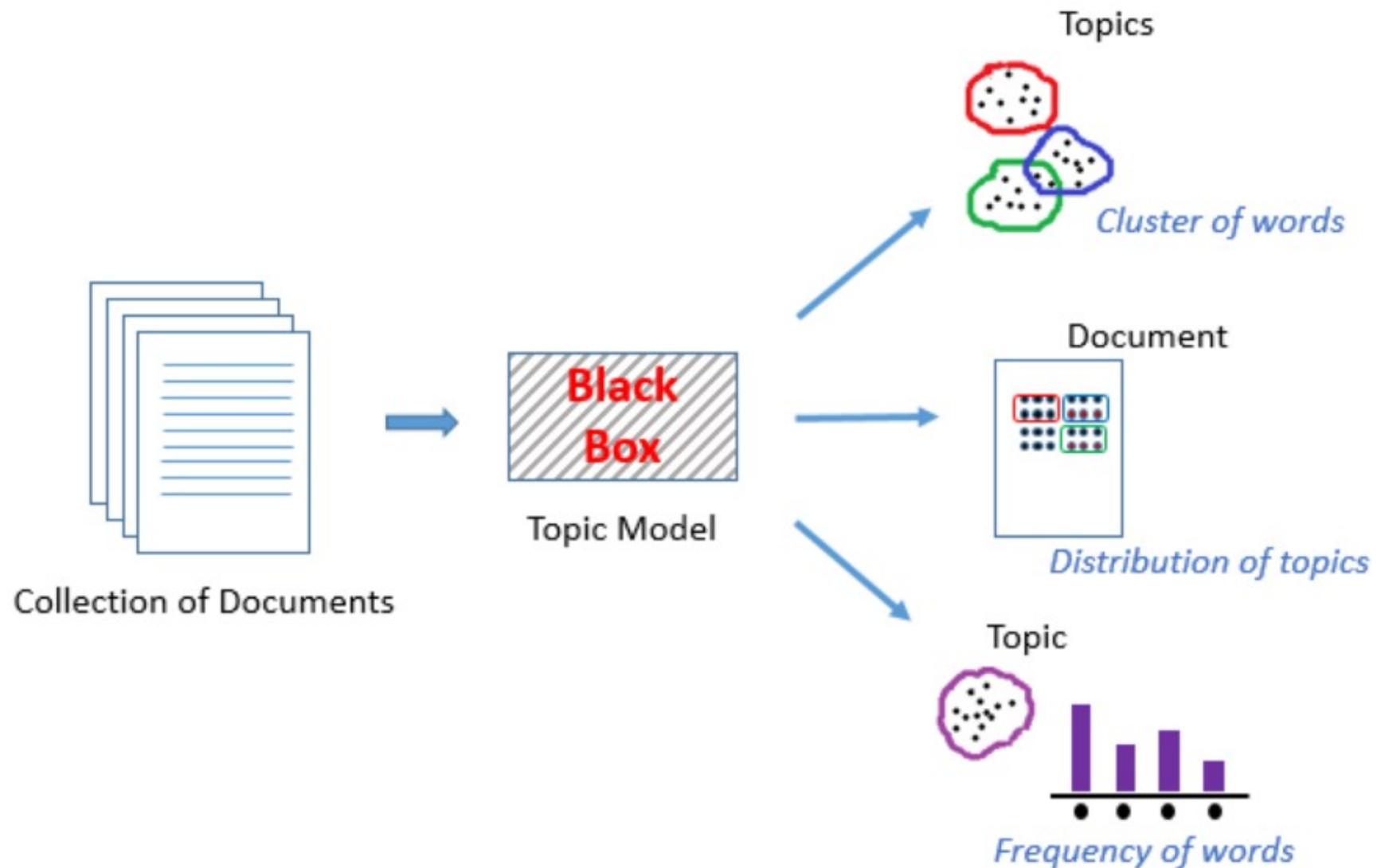
Modelagem de tópicos (MT)

Note, então, que estamos tratando um problema de **agrupamento de textos,** com a variante de que:

Documentos e palavras são agrupadas simultáneamente.

Existem diferentes métodos para modelagem de tópicos, as duas abordagens mais conhecidas são:

- **PLSA:** Probabilistic Latent Semantic Analysis (1999)
- **LDA:** Latent Dirichlet Allocation (2003)

(1) Trabalho pioneiro de Hofmann

HOFMANN, Thomas.

Probabilistic latent semantic analysis.

In: Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence.

Morgan Kaufmann Publishers Inc., 1999. p. 289-296.

(2) Trabalho mais popular de Blei *et al.*

Latent Dirichlet Allocation

David M. Blei   BLEI@CS.BERKELEY.EDU
Computer Science Division
University of California
Berkeley, CA 94720, USA

Andrew Y. Ng   ANG@CS.STANFORD.EDU
Computer Science Department
Stanford University
Stanford, CA 94305, USA

Michael I. Jordan   JORDAN@CS.BERKELEY.EDU
Computer Science Division and Department of Statistics
University of California
Berkeley, CA 94720, USA

Editor: John Lafferty

Abstract

We describe *latent Dirichlet allocation* (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document. We present efficient approximate inference techniques based on variational methods and an EM algorithm for empirical Bayes parameter estimation. We report results in document modeling, text classification, and collaborative filtering, comparing to a mixture of unigrams model and the probabilistic LSI model.

1. Introduction

In this paper we consider the problem of modeling text corpora and other collections of discrete data. The goal is to find short descriptions of the members of a collection that enable efficient processing of large collections while preserving the essential statistical relationships that are useful for basic tasks such as classification, novelty detection, summarization, and similarity and relevance judgments.

Significant progress has been made on this problem by researchers in the field of information retrieval (IR) (Baeza-Yates and Ribeiro-Neto, 1999). The basic methodology proposed by IR researchers for text corpora—a methodology successfully deployed in modern Internet search engines—reduces each document in the corpus to a vector of real numbers, each of which represents ratios of counts. In the popular *tf-idf* scheme (Salton and McGill, 1983), a basic vocabulary of "words" or "terms" is chosen, and, for each document in the corpus, a count is formed of the number of occurrences of each word. After suitable normalization, this term frequency count is compared to an inverse document frequency count, which measures the number of occurrences of a

BLEI, David M.; NG, Andrew Y.; JORDAN, Michael I.

Latent dirichlet allocation.

Journal of machine Learning research

v. 3, n. Jan, p. 993-1022, 2003.

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Figure 8: An example article from the AP corpus. Each color codes a different factor from which the word is putatively generated.

Modelos generativos e LDA

Modelos generativos para textos
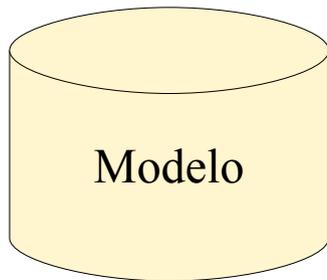
Modelo → Geração do documento → Documento

Modelos generativos para textos

Modelo

→
←

Estimação / inferência

BOW
A : 3677
Uma: 3400
Dom: 1000
Casmurro: 693

Documento

Modelos generativos para textos

Modelo

Modelo

Modelo

Modelo

Modelo

Mixture model!

Como inferir o modelo?
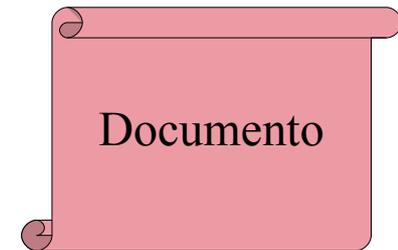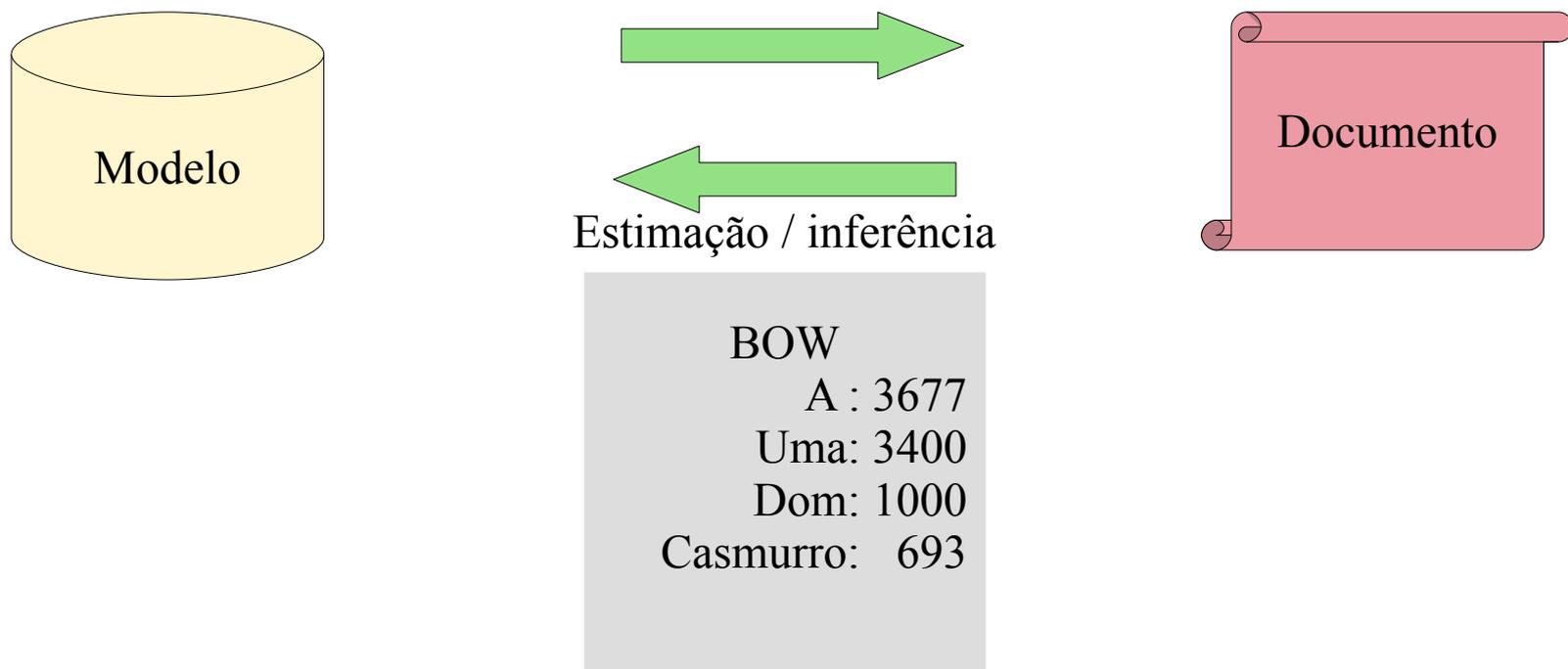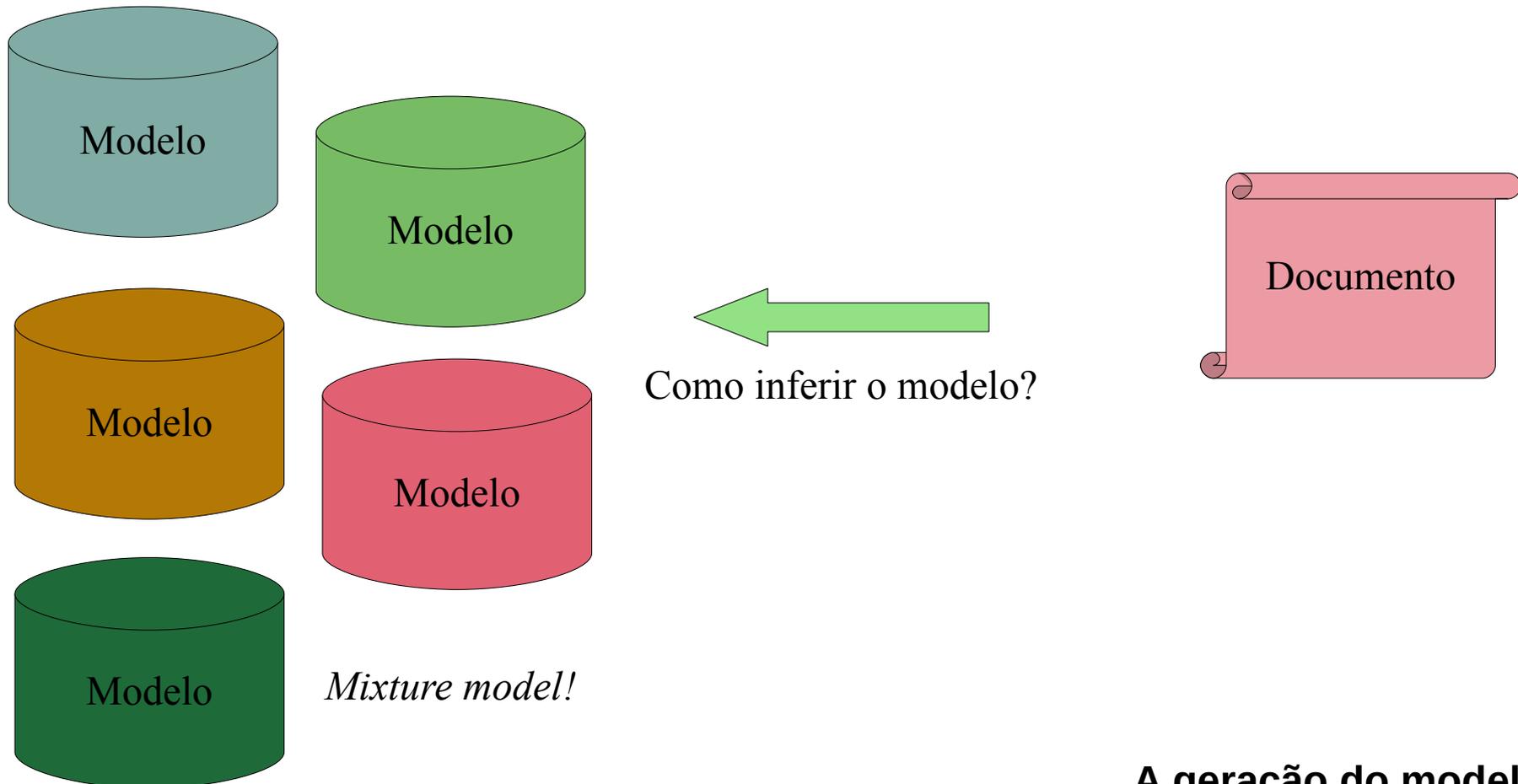
Documento

A geração do modelo é uma tarefa difícil.

LDA (Latent Dirichlet Allocation)

Documento	Tópico A	Tópico C	Tópico B
Tópico A Tópico B Tópico C	Palavra 1 (20%) Palavra 2 (13%) Palavra 3 (10%)	Palavra 1 (16%) Palavra 2 (02%) Palavra 3 (04%)	Palavra 1 (25%) Palavra 2 (11%) Palavra 3 (08%)
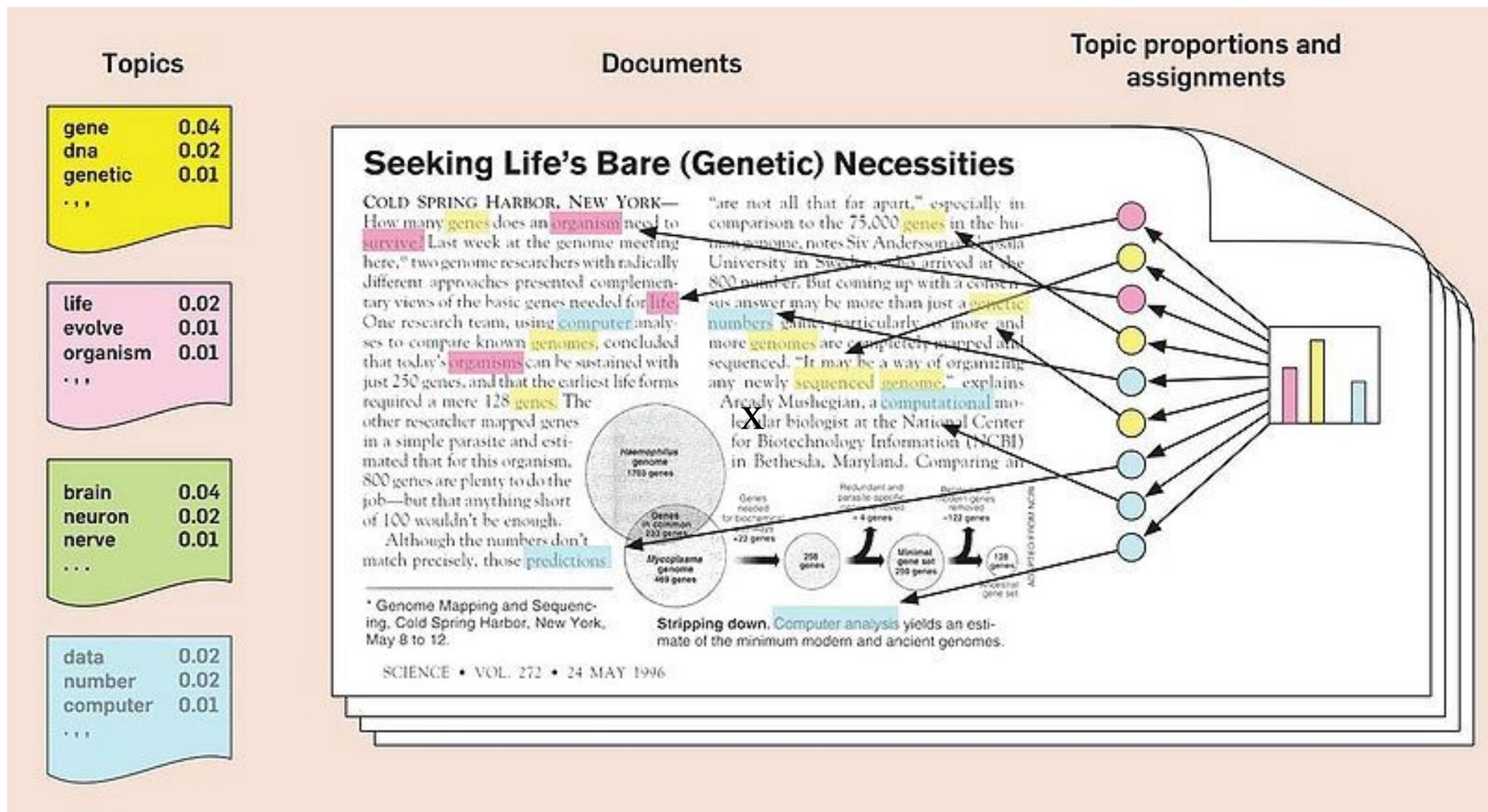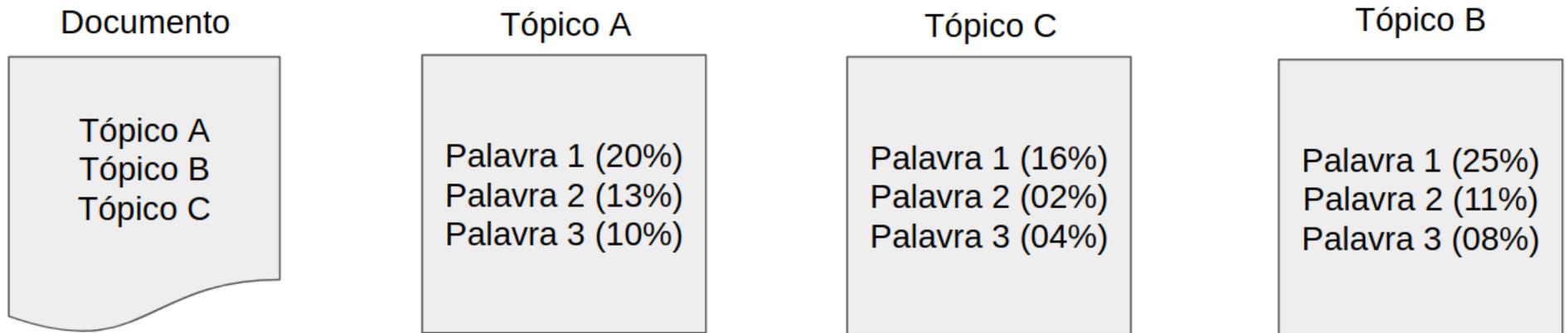
O problema aqui é identificar qual tópico é associado a um documento

Atribuir aleatoriamente os tópicos às palavras

(1)	(2)	(3)	(4)	(5)
A B C	C E D	A A B	E E B	E A E
B D E	D A D	C A B	B A E	B E B
A C C	A B E	B D E	D E A	D A E
C B A	E C A	E A B	C E E	A D B
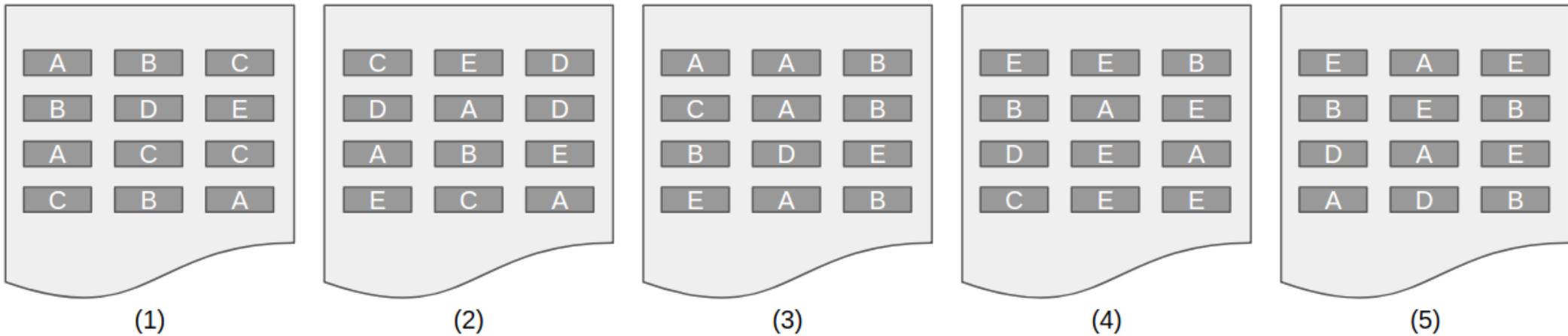
Na figura temos 5 documentos (1, 2, 3, 4 e 5) para serem classificados em 3 tópicos (vermelho, azul e verde).

Calcular P(T | D) e P(W | T)
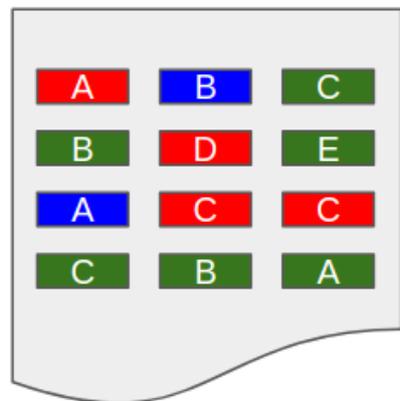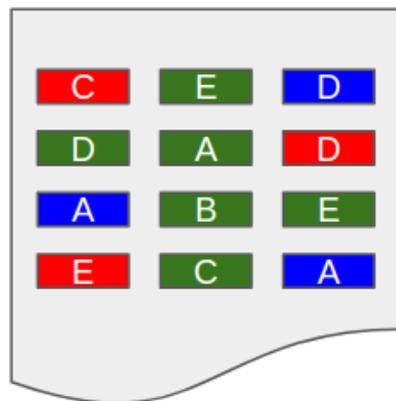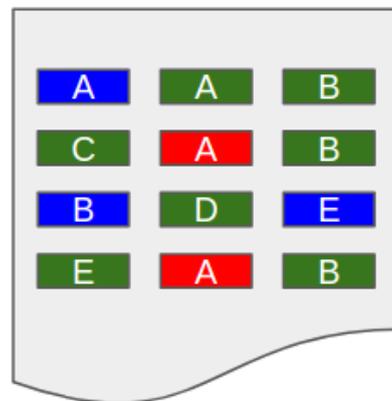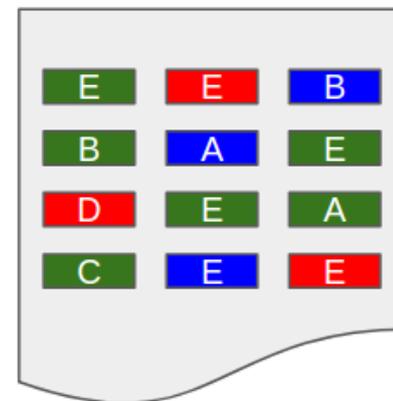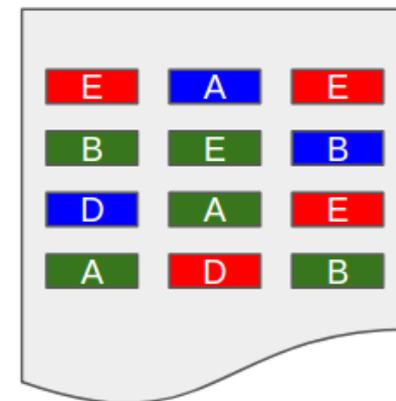
Aqui os tópicos (cores) foram atribuídos às palavras de forma aleatória.

A (0,05); B(0,00); C(0,05); D(0,07); E(0,10)

A (0,10); B(0,07); C(0,00); D(0,03); E(0,03)

A (0,10); B(0,15); C(0,08); D(0,03); E(0,13)

(1)
- 🔴 0,333
- 🔵 0,167
- 🟢 0,500

A	B	C
B	D	E
A	C	C
C	B	A

(2)
- 🔴 0,250
- 🔵 0,250
- 🟢 0,500

C	E	D
D	A	D
A	B	E
E	C	A

(3)
- 🔴 0,167
- 🔵 0,250
- 🟢 0,583

A	A	B
C	A	B
B	D	E
E	A	B

(4)
- 🔴 0,250
- 🔵 0,250
- 🟢 0,667

E	E	B
B	A	E
D	E	A
C	E	E

(5)
- 🔴 0,333
- 🔵 0,250
- 🟢 0,417

E	A	E
B	E	B
D	A	E
A	D	B
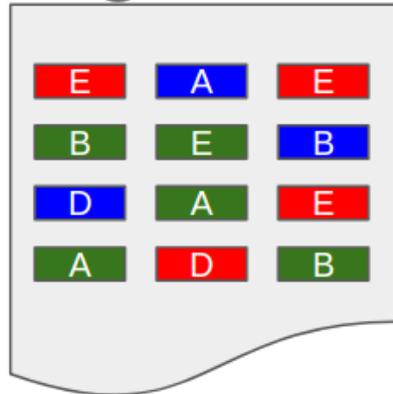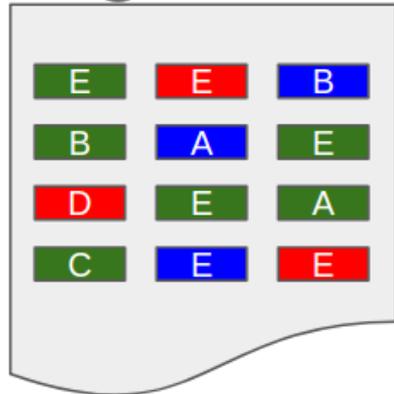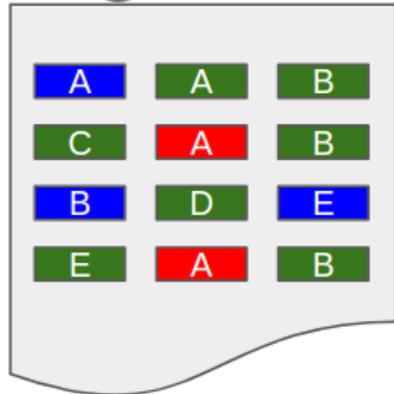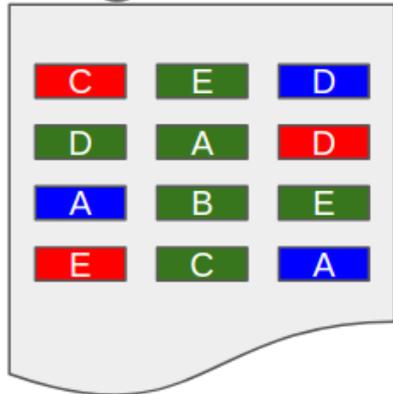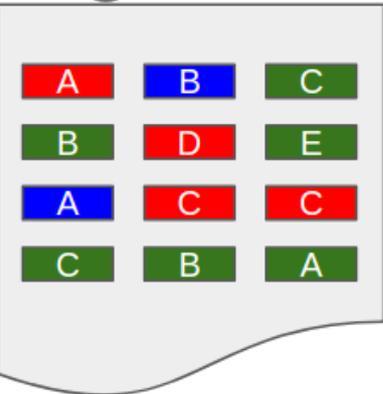
Atualizar o tópico (T) de cada palavra (W) com P(T | D) x P(W | T)

O cálculo de P(T|D) é a probabilidade do tópico no documento, no documento 1 o tópico vermelho é calculado da seguinte forma: 4 palavras vermelhas / 12 palavras (total) = 0,333.
O tópicos azul é: 2 palavras azuis / 12 palavras = 0,167, e assim por diante em todos os documentos.
O cálculo de P(W|T) é a probabilidade da palavra no tópico, para o tópico vermelho temos que: a palavra A está associada com o tópico vermelho 3 vezes em todos os documentos, assim: 3 palavras A com tópico vermelho / 60 palavras = 0,05.
A palavra B não está associada ao tópico vermelho, assim: 0 palavras B com tópico vermelho / 60 palavras = 0,00.

A atualização dos tópicos é feita palavra por palavra.

No exemplo a palavra A, que está associada ao tópico vermelho, tem os coeficientes do documento 1 multiplicados pelos coeficientes da palavra A nos tópicos.

A atualização do tópico na palavra é feita pelo tópico que tiver o maior resultado, no caso do exemplo: a palavra A foi atualizada com o tópico verde.

A (0,03); B(0,00); C(0,05); D(0,07); E(0,10)

A (0,10); B(0,07); C(0,00); D(0,03); E(0,03)

A (0,12); B(0,15); C(0,08); D(0,03); E(0,13)

(1)
0,250
0,167
0,583

A	B	C
B	D	E
A	C	C
C	B	A

(2)
0,250
0,250
0,500

C	E	D
D	A	D
A	B	E
E	C	A

(3)
0,167
0,250
0,583

A	A	B
C	A	B
B	D	E
E	A	B

(4)
0,250
0,250
0,667

E	E	B
B	A	E
D	E	A
C	E	E

(5)
0,333
0,250
0,417

E	A	E
B	E	B
D	A	E
A	D	B
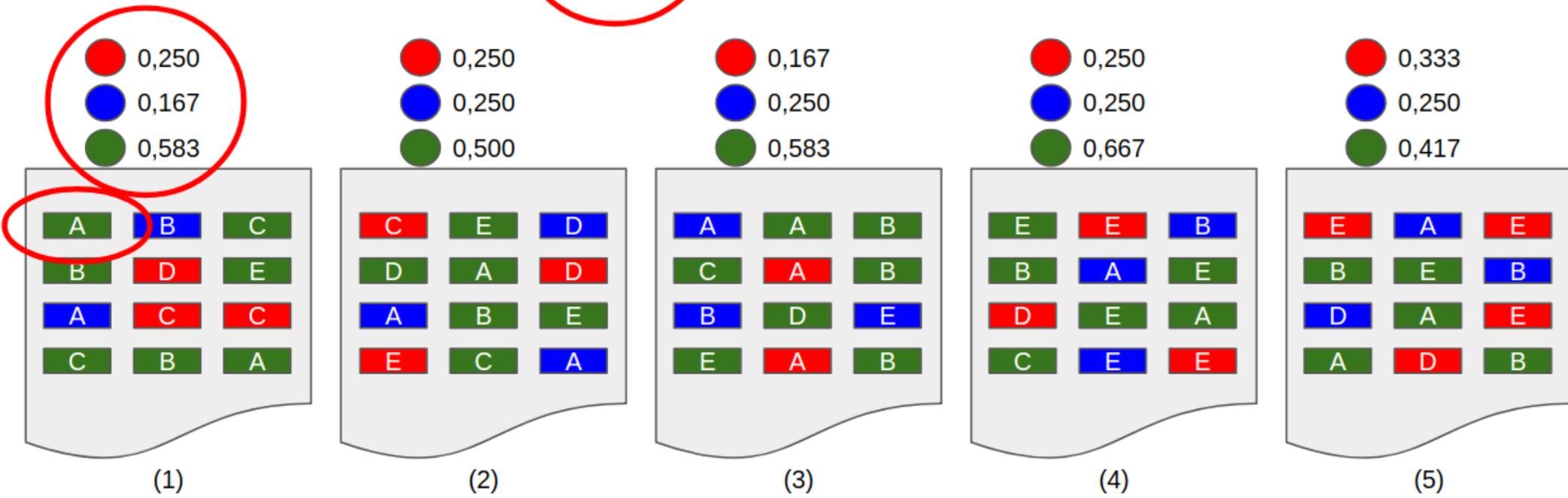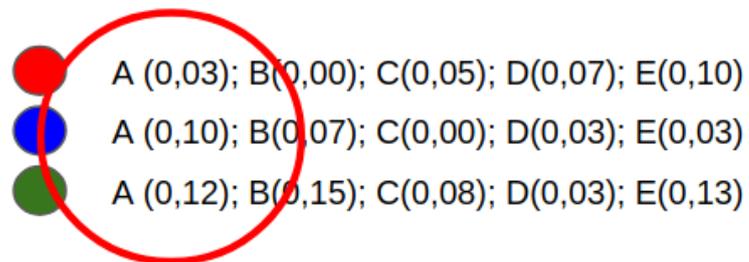
Assim, após a atualização do tópico verde na palavra A, os cálculos dos coeficientes dos documentos devem ser refeitos.

Também devem ser recalculados os coeficientes da palavra A nos tópicos.
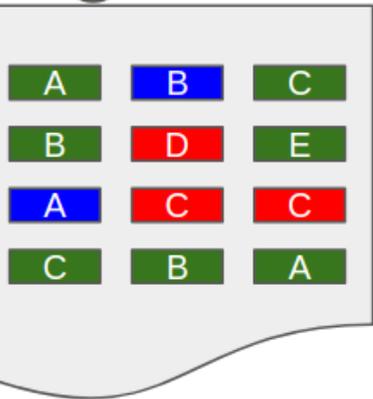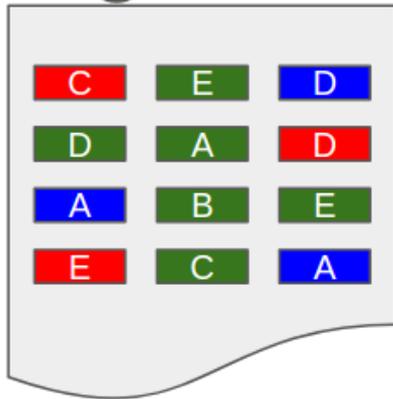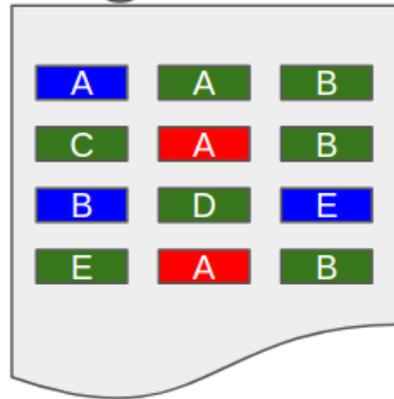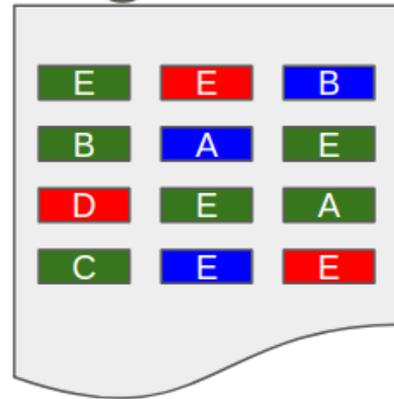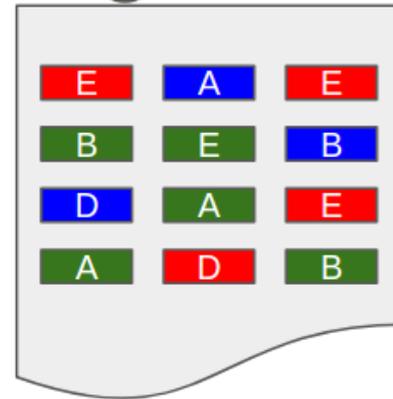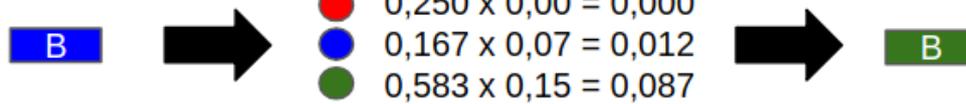
Esse exemplo é repetido para a palavra B, que no caso teve seu tópico alterado do azul para o verde.

A (0,03); B(0,00); C(0,05); D(0,07); E(0,10)
A (0,10); B(0,05); C(0,00); D(0,03); E(0,03)
A (0,12); B(0,17); C(0,08); D(0,03); E(0,13)

(1)	(2)	(3)	(4)	(5)
0,250	0,250	0,167	0,250	0,333
0,083	0,250	0,250	0,250	0,250
0,667	0,500	0,583	0,667	0,417

(1)
A B C
B D E
A C C
C B A

(2)
C E D
D A D
A B E
E C A

(3)
A A B
C A B
B D E
E A B

(4)
E E B
B A E
D E A
C E E

(5)
E A E
B E B
D A E
A D B
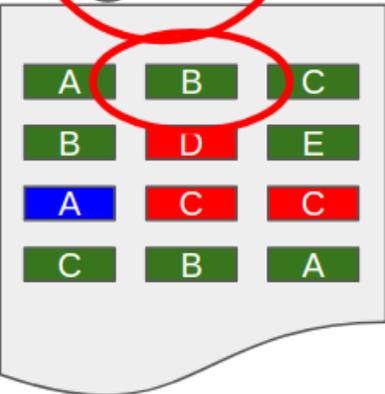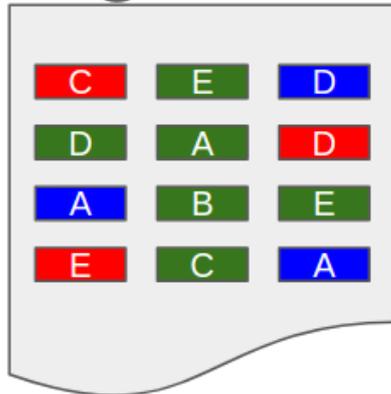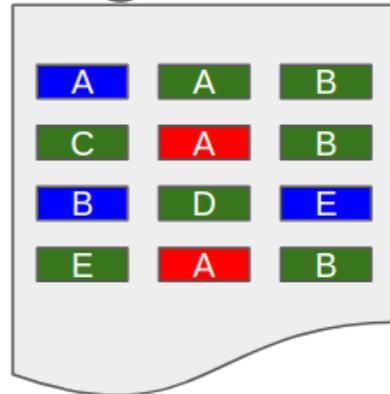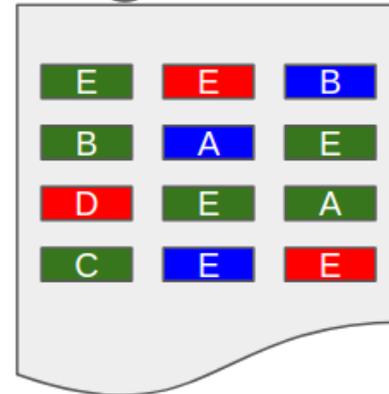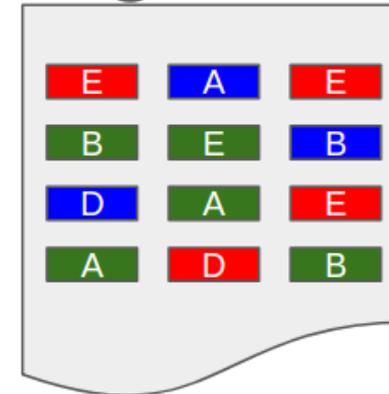
Aqui novamente o recálculo dos coeficientes do documento e do tópico

29

Exemplo (simples)

Suponha termos 5 documentos (com poucas palavras)

1) I like to eat broccoli and bananas.
2) I ate a banana and spinach smoothie for breakfast.
3) Chinchillas and kittens are cute.
4) My sister adopted a kitten yesterday.
5) **Look at this cute hamster munching on a piece of broccoli.**

Exemplo (texto padronizado)

Sem stop words, e após aplicação de um stemmer:

1) like, broccoli, banana
2) banana, spinach, smoothi, breakfast
3) chinchilla, kitten, cute
4) sister, adopt, kitten, yesterday
5) **look, cute, hamster, munch, piec, broccoli**

Exemplo (texto padronizado)

1) Topic: 0.117*"broccoli" + 0.116*"banana" + 0.076*"cute" + 0.072*"look" + 0.071*"hamster"

2) Topic: 0.117*"broccoli" + 0.116*"banana" + 0.076*"cute" + 0.072*"look" + 0.071*"hamster"

3) Topic: 0.158*"kitten" + 0.095*"sister" + 0.095*"yesterday" + 0.095*"adopt" + 0.094*"chinchilla"

4) Topic: 0.158*"kitten" + 0.095*"sister" + 0.095*"yesterday" + 0.095*"adopt" + 0.094*"chinchilla"

5) Topic: 0.117*"broccoli" + 0.116*"banana" + 0.076*"cute" + 0.072*"look" + 0.071*"hamster"

Redes de co-ocorrência aplicadas para a caracterização de obras literárias

Universidade Federal do ABC

Caracterização de obras literárias usando redes de co-ocorrência

Relatório Final - Iniciação Científica - FAPESP

Processo 2017/26981-9

Aluna: Bruna Pereira Santos

Podemos classificar obras literárias?
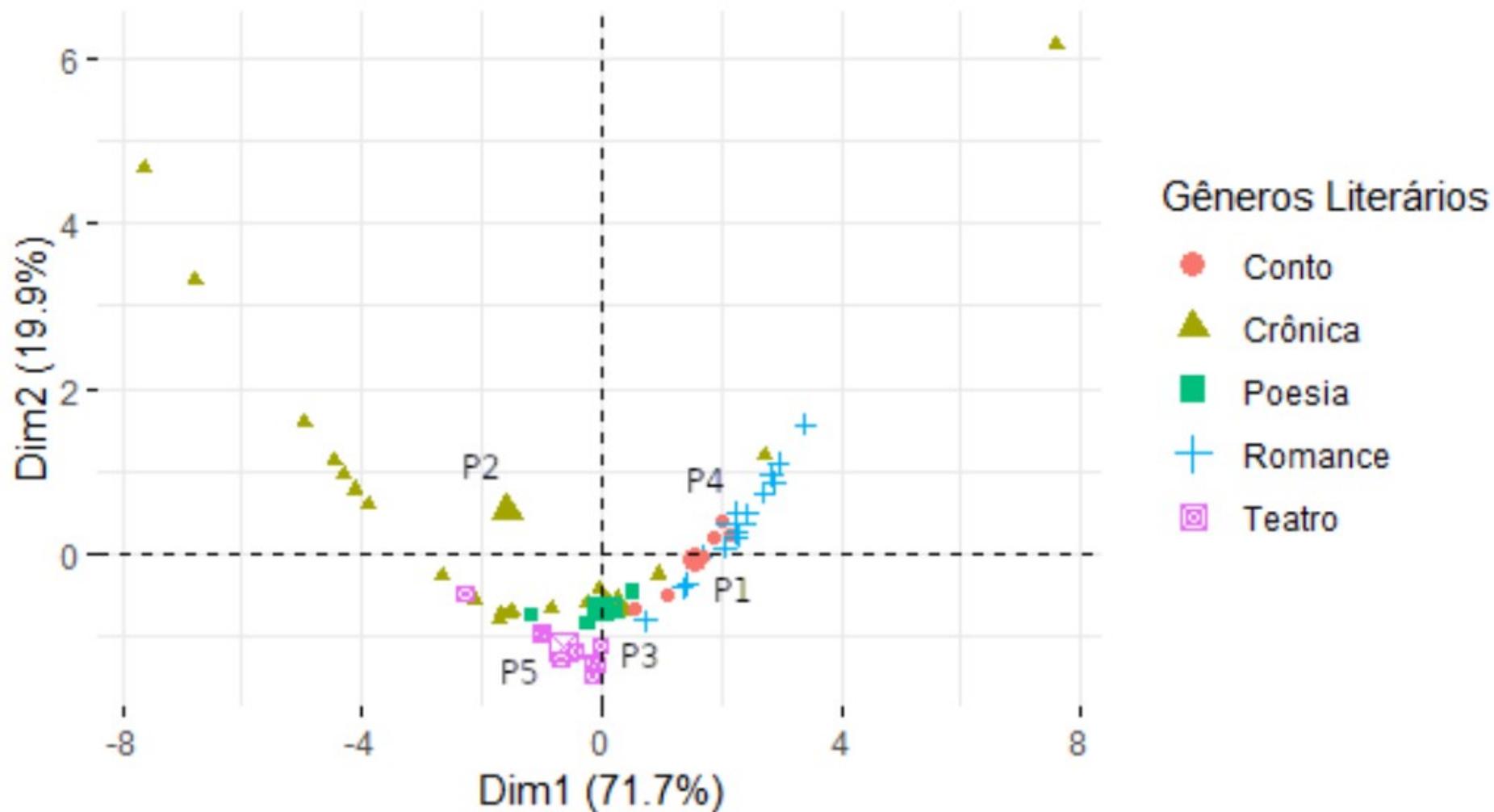
Objetivo do trabalho

Demonstrar, através de diferentes testes empíricos, que a **classificação de textos de obras literárias** pode ser realizada usando suas características estruturais (padrões topológicos) extraídas de redes de co-ocorrência textual.

Um grafo/rede de co-ocorrência

❖ **Frase original (Machado de Assis):**
"Talvez porque nenhuma tinha os olhos de ressaca, nem os de cigana oblíqua e dissimulada."

Um grafo/rede de co-ocorrência

❖ **Frase original (Machado de Assis):**
"Talvez porque nenhuma tinha os olhos de ressaca, nem os de cigana oblíqua e dissimulada."

❖ **Normalização (sem *stemming*):**
talvez porque nenhuma olhos ressaca cigana obliqua dissimulada

❖ ***Stemming:***
tal porqu nenhum olh ressac cigan obliqu dissimul

❖ **Rede de co-ocorrência (janela de conexão 2):**
tal porqu nenhum olh ressac cigan obliqu dissimul

❖ **Rede de co-ocorrência (janela de conexão 5):**
tal porqu nenhum olh ressac cigan obliqu dissimul

# Pipeline
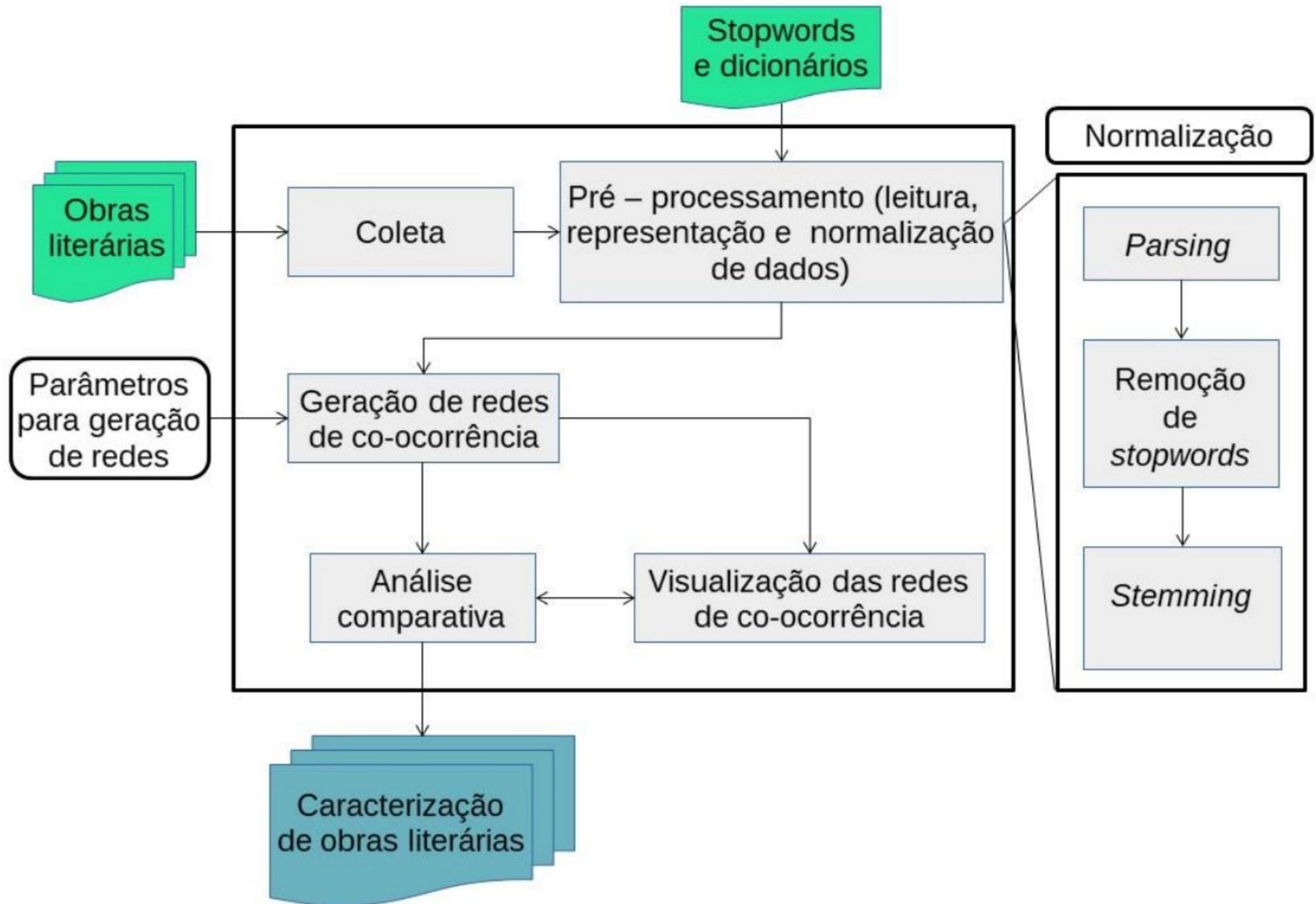
Características: Janela de tamanho 2

Obras/Atributos	Tipo de Grafo	Vértices (voca-bulário)	Arestas	Maior frequência entre palavras
Henrique Lobaerts sem *stemming*	Dirigido	100	110	2
Henrique Lobaerts com *stemming*	Dirigido	95	110	2
A semana sem *stemming*	Dirigido	23315	130546	94
A semana com *stemming*	Dirigido	10484	121179	96
A semana sem *stemming* (filtro aplicado)	Dirigido	227	185	94
A semana com *stemming* (filtro aplicado)	Dirigido	237	246	96

Características: Janela de tamanho 5

Obras/Atributos	Tipo de Grafo	Vértices (voca-bulário)	Arestas	Maior peso das arestas
Henrique Lobaerts sem *stemming*	Dirigido	100	436	2
Henrique Lobaerts com *stemming*	Dirigido	95	432	2
A semana sem *stemming*	Dirigido	23315	517191	100
A semana com *stemming*	Dirigido	10484	445152	109
A semana sem *stemming* (filtro aplicado)	Dirigido	279	358	100
A semana com *stemming* (filtro aplicado)	Dirigido	339	1018	109

(a) Sem aplicação do *stemming*

(b) Com aplicação do *stemming*

Obra "A Semana": Janela de conexão 5.
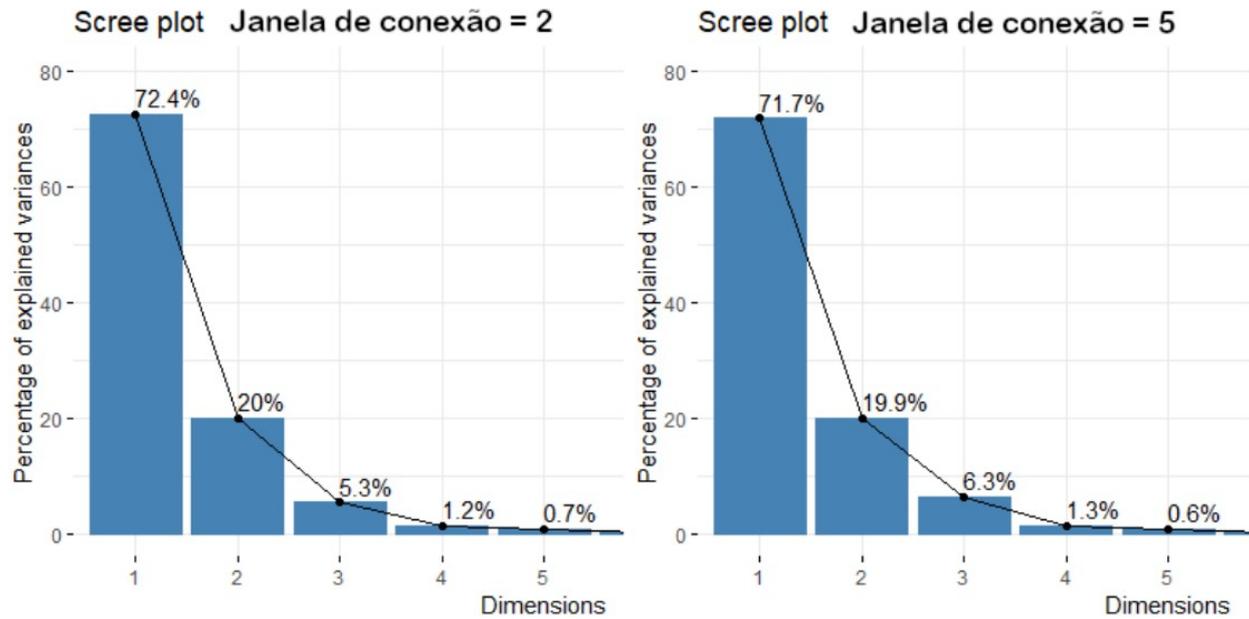
(a)
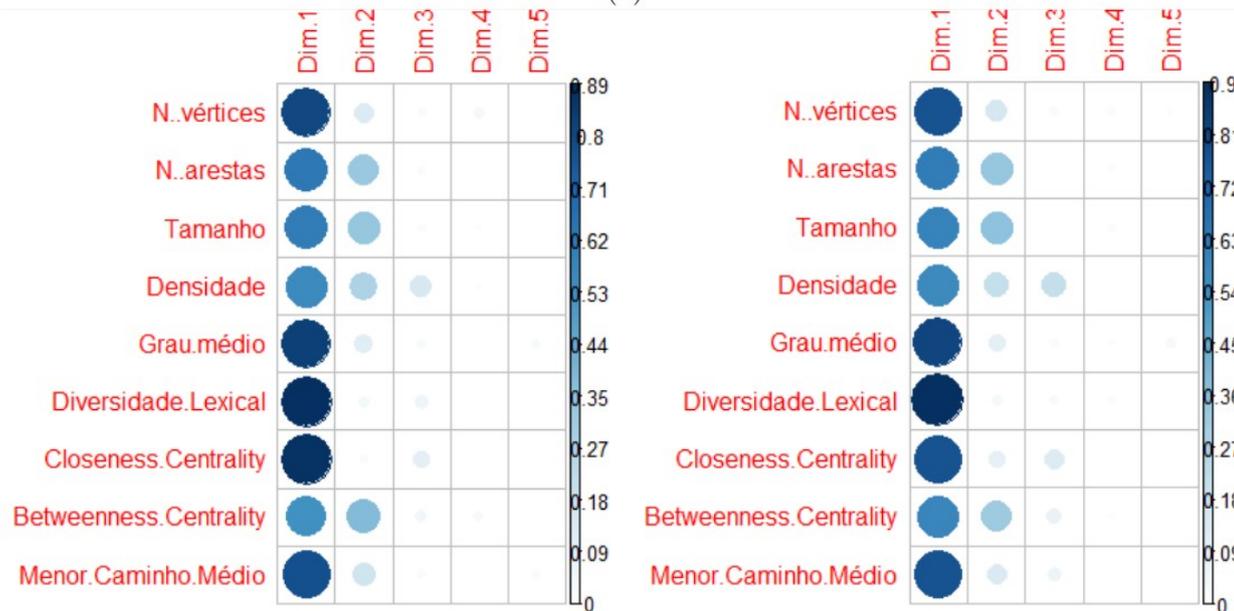
(b)

Janela de conexão 5
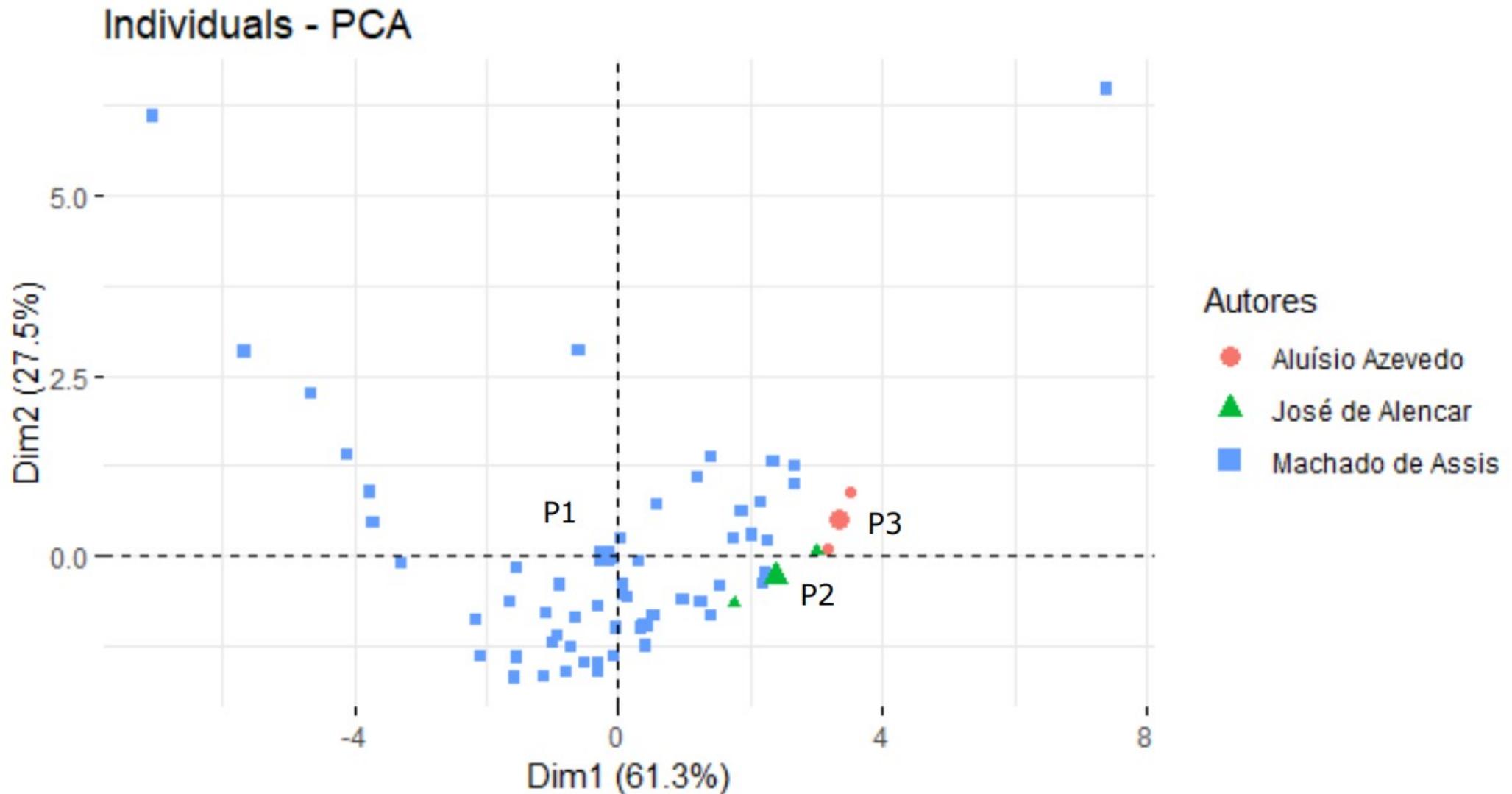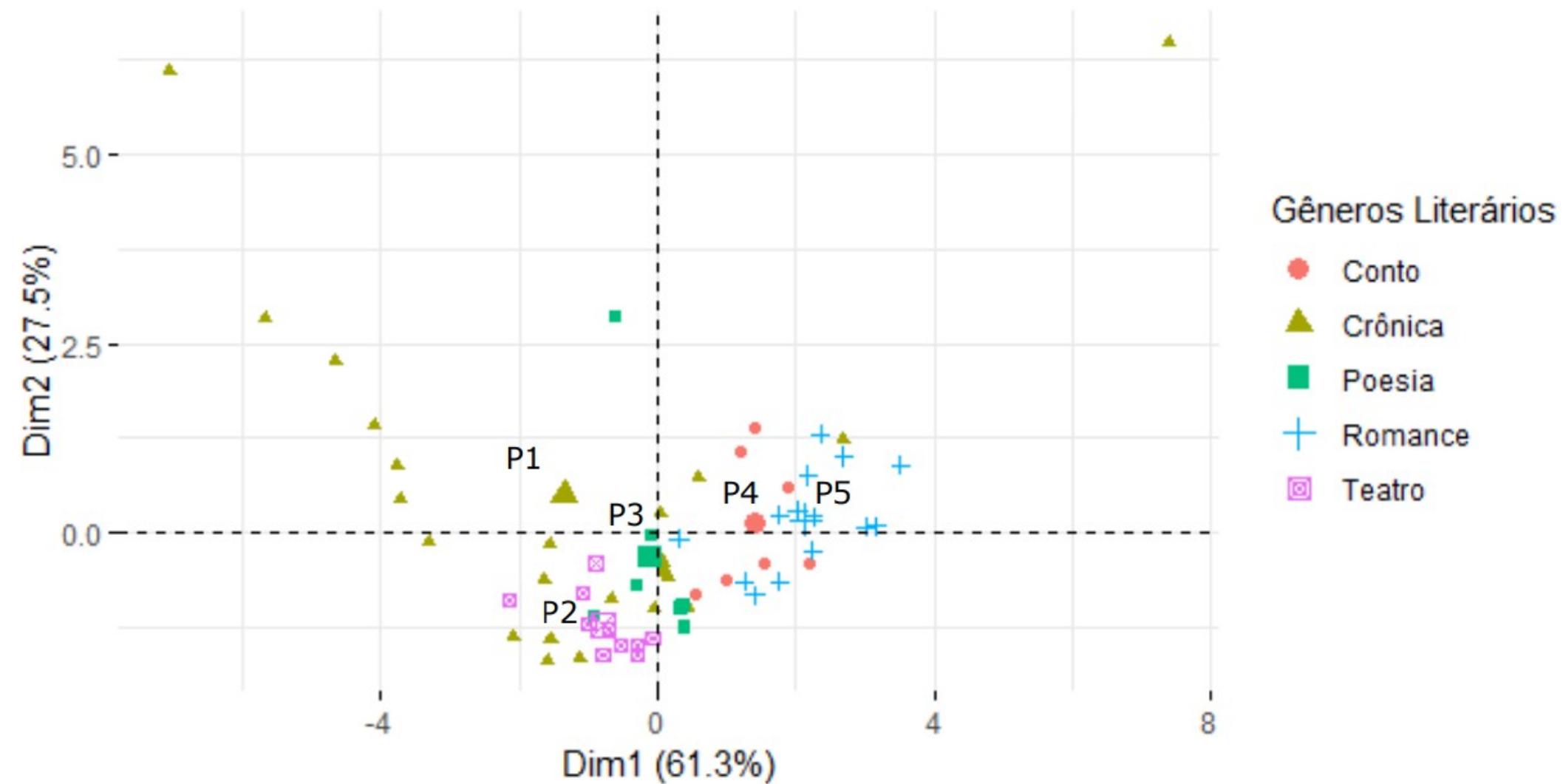
Comentários finais...

Com a visualização das redes de co-ocorrência é possível realizar uma análise mais qualitativa das obras literárias, sendo evidenciadas palavras de maior relevância, o que torna mais fácil detectar conceitos relacionados com a temática de cada texto.

- A partir das métricas obtidas das redes de co-ocorrência, é possível realizar uma análise mais quantitativa, o que permitiu buscar padrões de similaridade e comparar autoria e genêro literário.

- A técnica utilizada foi a de análise de componentes principais

sendo que, para ambas janelas de conexão, obtivemos uma porcentagem de explicação da variância de cerca de 90% com duas dimensões.